Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Web Footprints of Firms: Using Online Isomorphism for Competitor Identification

Gautam Pant, Olivia R. L. Sheng

To cite this article:
Gautam Pant, Olivia R. L. Sheng (2015) Web Footprints of Firms: Using Online Isomorphism for Competitor Identification. Information Systems Research 26(1):188-209. https://doi.org/10.1287/isre.2014.0563

*Information Systems Research*
Vol. 26, No. 1, March 2015, pp. 188–209
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)
http://dx.doi.org/10.1287/isre.2014.0563

# Web Footprints of Firms: Using Online Isomorphism for Competitor Identification

Gautam Pant
Department of Management Sciences, University of Iowa, Iowa City, Iowa 52242, gautam-pant@uiowa.edu

Olivia R. L. Sheng
Department of Operations and Information Systems, University of Utah, Salt Lake City, Utah 84112,
olivia.sheng@business.utah.edu

*Competitive isomorphism* refers to the phenomenon of competing firms becoming similar as they mimic each other under common market forces. With the growing presence of firms as well as their consumers and suppliers on the Web, we discover a parallel phenomenon of *online isomorphism* wherein the *Web footprints* of competing firms are found to overlap. We propose new online metrics based on the content, in-links, and out-links of firms' websites to measure the presence of online isomorphism as well as uncover its utility in predicting competitor relationships. Through rigorous analysis involving more than 2,600 firms, we find that predictive models for competitor identification based on online metrics are largely superior to those using offline data such as Standard Industrial Classification codes and market values of firms. In addition, combining online and offline metrics can boost the predictive performance. We also find that such models are valuable for identifying nuances of competitor relationships such as asymmetry and the role of industrial divisions. Furthermore, the suggested predictive models can effectively rank firms in an industrial division by their likelihood of being competitors to a focal firm as well as identify new future competitors, thus adding to a portfolio of evidence indicating their utility for managers and analysts.

*Keywords*: isomorphism; competitor identification; Web metrics; predictive models
*History*: Vijay Mookerjee, Senior Editor; Eric Zhang, Associate Editor. This paper was received on October 11, 2011, and was with the authors 22 months for 3 revisions. Published online in *Articles in Advance* February 25, 2015.

## 1. Introduction

*Isomorphism*, or the tendency of firms in an organizational field[1] to become similar, has been widely described in management and sociology literature (Hannan and Freeman 1977, DiMaggio and Powell 1983, Zander and Kogut 1995, Mizruchi and Fein 1999). Whereas different processes (DiMaggio and Powell 1983, Mizruchi and Fein 1999) may lead to firms becoming similar, isomorphism among firms operating in the same competitive markets can spontaneously appear in free markets without coercion. Hannan and Freeman (1977) provided a population ecology model for isomorphism that ensues from competition. The processes that lead to isomorphism often transpire in unison and are hard to distinguish from one another (DiMaggio and Powell 1983). However the overall effect of making competing firms more similar than noncompeting firms remains the same.[2] This general notion of isomorphism of competing firms motivates us to explore whether a parallel phenomenon of online isomorphism exists in the context of Web footprints of these firms. As the growing Web presence of firms, their clients/consumers, and their suppliers continues to augment firms' offline activities, we conjecture that the Web text and linkages related to competing firms should indicate a significant overlap. The presence of online isomorphism would provide a strong basis for designing new methods for identifying competitive threats that rely on overlapping Web footprints of firms.

The importance of identifying competitors and avoiding "competitive blind spots" (Zajac and Bazerman 1991) has been well documented (Walker et al. 2005). Competitor identification is a necessary precursor to competitor analysis and strategy (Bergen and Peteraf 2002). The significance of competitor identification is similar to that of requirement identification in software engineering, i.e., all of the subsequent analysis stands on the foundation of competitor identification. Previous literature describes difficulties, both from cognitive and procedural standpoints, in identifying competitors (Porac and Thomas 1990, Bergen and Peteraf 2002). Several

[1] An *organizational field* is a recognized area of institutional life that includes firms that produce similar services and products (DiMaggio and Powell 1983).

[2] We note that isomorphism does not imply that competing firms are identical. In fact, firms will retain their uniqueness and yet be more similar to competitors than noncompetitors.

frameworks for competitor identification that require manual efforts from managers have been suggested (Chen 1996, Bergen and Peteraf 2002). Clark and Montgomery (1999) describe the process of managerial identification of competitors as a manual categorization (i.e., competitors or noncompetitors) based on a pairwise similarity analysis of focal and target firms. The focal firm is the one for which the manager is seeking competitors, and the target firms define a space of potential competitors. There are two broad contexts in which competitor identification is needed: (1) when a manager is trying to identify competitors of the firm that she works for and (2) when a manager is trying to identify competitors of an external firm or firms. In the first scenario, the manager is expected to be well informed about the focal firm. In other words, she has a clear picture of one side of the story (i.e., the focal firm), but she may need to do (costly) research on the target firms for the categorization process. In the second scenario, the manager or analyst is an outsider without ready access to all of the internal (e.g., resources) and external (e.g., markets) information about the focal and target firms. This may be the case, for example, when a manager is looking for competitors of the firms in her firm's supply network or a financial analyst is seeking competitors of firms in a large portfolio (Ma et al. 2011). We note that in the second scenario there may be more than one focal firm of interest. Both scenarios may suffer from "managerial myopia" (Bergen and Peteraf 2002) in recognizing competitors. However, because of a lack of information about the focal firms and the difficulties in scaling a manual process over a large number of focal firms, the second scenario is much more challenging and in greater need of automated tools. In general, the competitor identification problem can be combinatorial in the number of focal and target firms, and the nonstationary nature of the relationships requires constant monitoring over time. We present a use case in Appendix A of the online companion (available as supplemental material at http://dx.doi.org/10.1287/isre.2014.0563) that illustrates a potential tool that can allow a user to identify competitive threats as well as drill down into similarities in online presence of multiple focal and target firms. The use case also illustrates the possibility of identifying future competitors. Such a tool would benefit both of the above-mentioned scenarios.

Using over 2,600 firms and their competitors from the Russell 3000 index (Russell Investments 2009), we ask the following questions: (1) Do competing firms display online isomorphism? (2) To what extent are competitive relationships discernible from metrics of online isomorphism? We begin by suggesting carefully crafted Web metrics that require us to crawl hundreds of thousands of Web pages across firms in our data set. We further obtain a list of more than a million URLs of pages that link to websites of those firms. The new metrics that we suggest relate to firms' Web footprints as revealed by the content of their websites and the associated linkage structure (in-links and out-links). We suggest additional *control metrics* that are derived from previous works in text and Web mining for interfirm relationships. The control metrics leverage on billions of Web pages and millions of news stories indexed by search engines. Together these Web metrics measure the overlap between various types of Web footprints of a pair of firms and thus provide us with different perspectives of online isomorphism. We then present a systematic study that explores the signal (if any) contained in these metrics that is relevant to the problem of automatic competitor identification. More specifically, our study makes the following main contributions:

- We provide direct evidence to support the presence of online isomorphism by showing that Web footprints of competing firms have significantly greater overlap than the Web footprints of noncompeting firms.
- We propose three new Web metrics of online isomorphism based on firms' website content and linkage structure. These are used, along with metrics derived from literature that utilize online news and search engine results, to feed predictive models for competitor identification. Such an integration of diverse Web sources in automating competitor identification has not been explored previously. We find that the resulting predictive models are effective. We also tease out and highlight the predictive value of the newly crafted metrics.
- We benchmark the performance of the predictive models for competitor identification using online information against those using offline information about firms. This type of benchmarking, although practically essential and useful, is not seen in the previous works on automated competitor identification. We also test models that use both online and offline information and find that the combination of the two sources outperforms the individual sources.
- We explore predictive models of more nuanced relationships such as competition among firms in the same industry and symmetric versus asymmetric competition.
- We also consider the problem of ranking all firms within an industry by their likelihood of being competitors to a given focal firm. This problem represents an exhaustive intraindustry exploration for competitors.
- We uncover the utility of predictive models based on online isomorphism in not only identifying contemporary competitors but also new future competitors.

- We present an evaluation framework where the data sets and the methods are controlled so that fair and direct comparisons can be made between various metrics. For example, we present results with data sets that have different ratios (and hence prior probabilities) of competitors and noncompetitors. These different data sets represent various levels of information deficit in competitor identification. The large number of firms covered, the variety of online and offline metrics used, and the tight control over methods and data allow for the most rigorous evaluation framework for the competitor identification problem in the literature.

The suggested models that use online metrics for competitor identification can be used for both public and private firms. We validate this approach on a large number of public firms since competitor data (necessary for validation) for public firms are relatively easier to obtain. However, nothing restricts the suggested approach to public firms, since private firms have substantial Web footprints as well.

Commercial firm profiling resources such as Hoovers and Mergent can also benefit from our approach since the suggested approach can complement their manual efforts as well (in fact, they fall into the second scenario described earlier, i.e., a manager looking for competitors of external firms). We note that Ma et al. (2011) have shown that the list of competitors provided by commercial data sources such as Hoovers is incomplete. The incompleteness of these data sources is not surprising given the manually intensive process of competitor identification and the dynamic nature of competitor relationships. However, despite its incompleteness, Hoovers remains an industry leader in providing such data, and hence we use it as a gold standard for building and evaluating our models throughout this study.

Given that competitor identification remains a largely manual effort and the dynamic nature of global competitive environments poses additional urgency for timely competitor identification, there is a strong need for alleviating the complexity of the problem through automated methods. Our study unravels the phenomenon of online isomorphism and provides a much needed systematic exploration of a variety of Web metrics for automatic competitor identification.

## 2. Related Work

Several works in strategy literature have discussed the need for the accurate identification of competitors and provided theoretical frameworks for that purpose (Chen 1996, Bergen and Peteraf 2002, Peteraf and Bergen 2003). Competitor identification is referred to as a classification process through which competitors of a focal firm are identified based on "relevant similarities" (Bergen and Peteraf 2002). Given the expected isomorphism between competing firms (Hannan and Freeman 1977), the process of competitor identification through pairwise analysis of similarities between focal and target firms (Clark and Montgomery 1999) is well founded. The unit of analysis is a pair of firms since competitor relationship is seen as a unique interaction between the pair (Chen 1996). Bergen and Peteraf (2002) and Chen (1996) suggested frameworks for the manual identification of competitors. The manual nature of these frameworks makes them very costly for competitor identification over a large number of focal and target firms, and over time.

A few papers in the text- and web-mining literature have explored the idea of finding relationships between firms using news articles. The earliest work in this direction was by Bernstein et al. (2002), where they created virtual links between firms if they were mentioned in the same piece of news. A firm that appeared in a large number of news stories with other firms had a large number of links. Using link analysis, centrality of firms was measured and it was found that the 30 most central firms in the computer industry included several Fortune 1000 firms. Bernstein et al. (2003) used the co-occurrence of stock tickers in news stories to identify connections between firms and utilize the resulting network to predict the industry membership of firms. More recently, Ma et al. (2011) used the co-occurrence of stock tickers of firms in news stories to create connections between firms and use properties of the resulting network to predict competitor relationships. All of these works use just one source of information (business news) and are based on the assumption that co-occurrence of firms in news stories indicates a potential relationship between the firms.

Bao et al. (2008) described a competitor-mining system based on the "observation" that competitors tend to co-occur in Web pages (and hence search engine results) more often than noncompetitor pairs. The authors provide no direct empirical support for this observation, but utilize this observation/hypothesis in the design of their competitor-mining system. Their proposed approach depends on just one data source (search engine results), and the evaluation is based on a very small set of firms ($< 100$). We feel that restricting web-based automatic identification of competitors (or other relationships) to a single data source (news or search engine) would limit the approach from gaining a broader view of firms' footprints on the Web, and hence limit its performance. In contrast to previous works, we propose predictive models that utilize a variety of Web resources that are publicly available. Although the resources that we use do not exhaust all possible sources of web-based information, they do represent a significant step forward, in terms of the breadth, compared to the literature. Moreover the various web-based metrics that we propose and use are carefully crafted to provide complementary information that can help the predictive models in triangulating the evidence. Also, we conduct our study with data collected for a large number of firms ($> 2{,}600$), providing greater validity and generalizability to our results. Finally, we benchmark our models based on online data with similar models based on offline data, which we believe is an essential and practically useful exercise that has been largely ignored in the related literature.

Some of the Web metrics and models used in this study are analogous to or derived from those discussed in the context of social networks analysis (Scott 1991), bibliometrics (Egghe and Rousseau 1990, Bar-Ilan 2008), information retrieval (Manning et al. 2008), and machine learning (Witten and Frank 2005). We will make connections to the relevant literature while discussing the metrics, models, and their evaluation.

More generally, there is recent literature on the use of text and Web mining for business intelligence and marketing research (e.g., Lee and Bradlow 2011, Zheng et al. 2012). These works describe complementary efforts that can benefit from the methods, such as the one proposed in this study, that can alleviate the combinatorial and temporal nature of monitoring competitive threats.

## 3. Online Isomorphism

Motivated by the general concept of isomorphism, we suggest, test, and exploit the analogous phenomenon of online isomorphism where we conjecture that the Web footprints of competing firms are more similar than Web footprints of noncompeting firms. Our online isomorphism conjecture is based on the general premise that the Web presence of firms, large and small, has enormously increased since the late 1990s (Heinze and Hu 2006). This wide adoption of the Web is making the firm-level Web content increasingly reflective of firms' activities and relationships. Hence, isomorphism between competing firms, previously observed in the offline context, is expected to be applicable to the online sphere as well. There are several different ways in which Web footprints of firms may materialize. Table 1 describes several types of Web footprints related to firms. We note that although social media platforms such as Twitter and Facebook are gaining popularity among firms, the adoption and use of these channels by firms across industries and firm sizes is not yet comparable to the widespread adoption and use of corporate websites. A recent survey by Gartner (2013) suggests that "the corporate website will not be displaced anytime soon by a brand's social media presence."[3] Given that our study involves a large number of firms (more than 2,600) from various industries, it is important for us to focus on the Web footprints that cover the largest subset of these firms. Hence, we primarily focus on websites of firms and their surrounding linkage structure to suggest metrics of online isomorphism. Identifying and using website-based footprints of firms for competitor identification is a unique contribution of our study. However, we do cover large amounts of other Web content (which may include social media) through the use of Yahoo Application Programming Interface (API) for news and general Web search.

**Table 1 Types of Web Footprints of Firms**

| Web footprint | Description |
|---|---|
| Firm's website | A website that is owned by the firm and provides a description of a firm's activities, products, and services to its various stakeholders; the firm's website may include links to other sites that are related to the firm |
| In-linking sites | These are websites (such as consumer blogs) that link to the firm's website potentially providing a perspective of various stakeholders of the firm |
| Online news | News stories available on the Web from a large number of news sources that mention the firm |
| Social media sites | These include popular platforms such as Twitter and Facebook that are being increasingly used by firms to communicate with various stakeholders |
| General Web pages | Web pages, in addition to those described above, such as blogs, reviews, etc., that are captured by general-purpose search engines |

Although we may expect offline isomorphism of competing firms to spill over to their Web footprints, one may ask, How does online isomorphism manifest itself? Why should online content related to competing firms be similar and in what way? We note that the focus of the current work is not on unraveling the processes that create online isomorphism, but rather on utilizing the observed similarities in online content (that we refer to as online isomorphism) to predict competitor relationships.[4] However, for a richer discussion, we present some potential explanations for the observed similarities in Web footprints of competing firms.

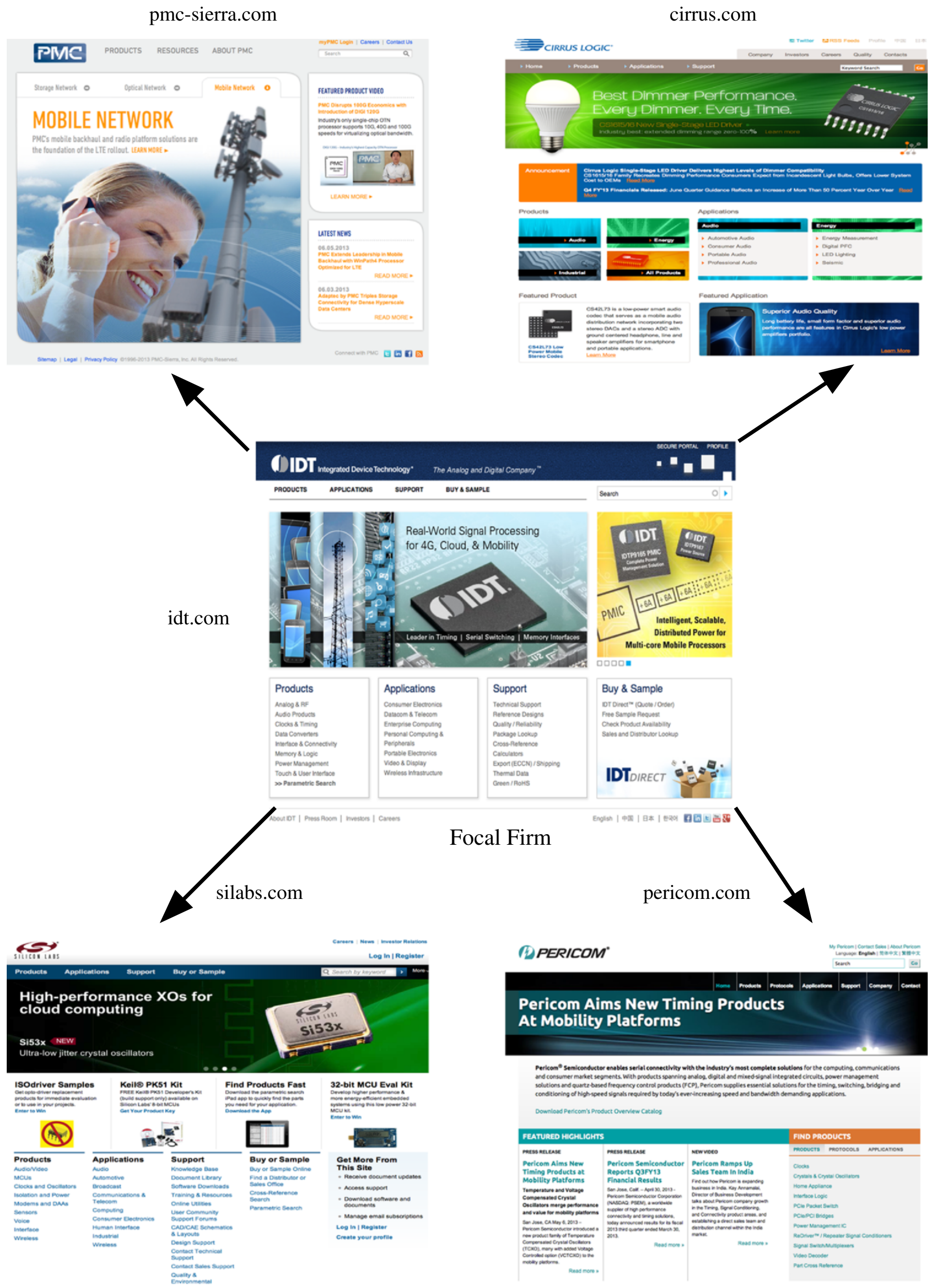
1. *Business communication*. Websites provide a popular medium for business communication, and firms of all sizes and industries use their websites to communicate with their various stakeholders (e.g., consumers, suppliers, investors, community, etc.) (Hill and White 2000, Esrocka and Leichtya 2000, Kent et al. 2003, Heinze and Hu 2006). While developing an institutional theory of organizational communication,

[3] http://www.gartner.com/newsroom/id/2368315.

[4] Predicting has been noted as an important complementary effort to providing explanations (Shmueli 2010).

Lammers and Barbour (2006) highlight the role of isomorphism in communication. Given that websites of competing firms are targeting similar stakeholders, we can expect their communication to be more similar than that of noncompeting firms. To illustrate, Figure 1 shows the home pages of a focal firm (Integrated Devices Technology, or IDT) and some of its competitors from our data set. Note that this example only shows the home pages of these firms, and yet we find many words such as "cloud," "mobile," "audio," "video," "clocks," and "telecom" that are common to the focal and one or more competitor websites. We expect even more common words appearing among these sites as we browse more pages on the sites. Similar to text, links provide important business information (Vaughan and Gao 2006). It has been observed previously that although competitors' websites rarely link to each other, they are often colinked (i.e., linked by/to the same third parties) (Pant and Menczer 2003, Vaughan and Gao 2006). This is to be expected since

**Figure 1** (Color online) Home Pages of a Focal Firm (idt.com) and Its Competitors


pmc-sierra.com
cirrus.com
MOBILE NETWORK
Best Dimmer Performance. Every Dimmer. Every Time.
idt.com
Real-World Signal Processing for 4G, Cloud, & Mobility
Intelligent, Scalable, Distributed Power for Multi-core Mobile Processors
Products
Applications
Support
Buy & Sample
Focal Firm
silabs.com
pericom.com
High-performance XOs for cloud computing
Si53x NEW
Ultra-low jitter crystal oscillators
Pericom Aims New Timing Products At Mobility Platforms

*Note.* The text on the home pages alone show several overlapping words.

the websites of competing firms are communicating with overlapping stakeholders and hence are likely to gain attention (through links) from similar third parties. Firms are using their websites as a communication medium, and since the target groups of these communications will overlap among competing firms, we can expect a consequent overlap in terms of both text and links among competitors' websites.

2. *Technology adoption*. The website of a firm is a type of information system (Albert et al. 2004) that is owned and maintained by the firm. It has been observed that competing firms tend to adopt similar technologies (Santos and Peffers 1998). In particular, Flanagin (2000) noted that "social pressures" among firms can explain, at least in part, the website adoption decisions. If a website is viewed as an information system, content is an important component of the system that is expected to be continuously monitored by stakeholders (including competitors) and updated by the firm. Given these observations we can expect similarities in the content of competitors' websites as they adopt similar web-based systems.

3. *Search cost*. Search engines serve as critical information brokers between the producers and consumers of online information. They greatly reduce the search cost of information consumers by presenting them with relevant content from information producers. The producers of online information, such as firms with websites, are interested in ranking highly for keywords that their stakeholders (e.g., customers) use to search for information. This has led to a search engine optimization (SEO) industry that is focused on modifying content of websites to improve their ranking on search engine results for various keywords. Popular SEO strategies include monitoring competitor websites for keywords and discovering keywords that target users are likely to search.[5] There are even tools designed to find competitor's profitable keywords.[6] Since websites of competing firms have overlapping target users, and, in addition, if they adopt a similar set of "successful" keywords, we can expect greater similarity in website content between competitors than noncompetitors.

In a Harvard Business School background note, Rivkin and Cullen (2008) discuss the use of company websites in manual industry analysis. In particular, they highlight the need to analyze the content of the company websites (how does a company describe its business on its site?) and also the links going out (out-links) from the company websites. For the latter, they note that the out-links of a firm's website can provide signals on the aspects of the external environment that the firm considers important (Rivkin and Cullen 2008). Hence, in practice, the importance and need for analyzing websites of firms for competitor identification has been recognized. However, these suggestions from practice are limited to manual analysis based on potentially anecdotal evidence. Our study fills in the gap by providing a systematic analysis of automated competitor identification through large-scale Web data. It does so by first empirically revealing the phenomenon of online isomorphism between competing firms, which provides a much needed framework for using Web-based signals for competitor identification.

## 4. Test Bed

Our study is based on firms in the Russell 3000 index. The firms in the index represent "approximately 98% of investible U.S. equity market" (Russell Investments 2009). The firms belong to different industries ("Services," "Mining," "Retail," etc.) and are of different sizes (large cap, mid cap, and small cap). For each of the firms listed in the index, we attempt to find the corresponding website URL from Yahoo! Finance. We also use the Hoovers API[7] to obtain a list of competitors for each of the firms. Since our analysis is restricted to Russell 3000 index firms, we consider only those competitors that are in the index as well. The firms for which competitors are obtained from Hoovers constitute the focal firms for our experiments (Chen 1996). Some of the Russell 3000 index firms are filtered out in the data collection process if Yahoo! Finance does not provide a corresponding URL or if no competitor firm (from the Russell 3000 index) is provided by Hoovers. Finally, we are able to identify a data set of 16,485 competitors (pairs) across 2,678 focal firms.

The U.S. Department of Labor (DOL) divides firms into 10 divisions using the Standard Industrial Classification (SIC) codes of the firms.[8] For the rest of this paper, we will refer to these standard divisions specified by the U.S. DOL as *industrial divisions*. Based on the U.S. DOL specification, as shown in Online Appendix B, we map firms to their industrial divisions. Using the 2,678 firms and their corresponding competitors, we create the following data sets for our experiments.

### 4.1. Any-Division Pairs

As noted above, we identified a data set of 16,485 competitors (pairs) across 2,678 focal firms. For each of the focal firms, we identify an equal number of random (noncompetitor) firms from the Russell 3000

[5] See http://www.wordstream.com/blog/ws/2012/10/03/keyword-content-marketing-faq#.

[6] See http://www.spyfu.com/?gclid=SCLGW_-C30LcCFa9eQgod5ioAeQ.

[7] See http://developer.hoovers.com/.

[8] See http://www.osha.gov/pls/imis/sic_manual.html.

index to create a data set of 16,485 noncompetitors (pairs). Note that we allow the firms in a competitor or noncompetitor pair to be from the same or different industrial divisions. This allows for the broadest coverage of competitor pairs. We will refer to this data set of 32,970 (16,485 × 2) pairs of firms as the any-division pairs (ADP) data set. We use this data set to study the performance of the predictive models with different Web metrics of online isomorphism. We do so while controlling for the effects of algorithms, skewness in class distributions (competitors versus noncompetitors), offline similarity metrics using SIC codes and market cap, and industrial divisions.

### 4.2. Same-Division Pairs

We wanted to explicitly study the role of industrial division in online isomorphism. For this purpose, starting from the ADP data set described above, we retain only noncompetitor pairs where the two firms (in a pair) are from the same industrial division. As a result, we obtain 3,561 pairs of same-division noncompetitors. We couple these with a random sample of same-division competitor pairs from the same ADP data set. Hence, we have 7,122 (3,561 × 2) within-industry pairs of firms in this filtered data set that we call the same-division pairs (SDP) data set. We use them to explore the differing empirical signals of online isomorphism between within-industry competitors and noncompetitors.

### 4.3. Exhaustive Same-Division Pairs for Holdout Firms

We randomly select 100 firms from our data set of 2,673 firms. We call these firms *holdout firms* to indicate that they represent firms for which no a priori information is available for building predictive models. For each of the holdout firms, we identify an exhaustive list of all firms in our data set that fall into the same industrial division (as the holdout firm). We then prepare pairs of firms by taking each holdout firm and matching it with a same-division firm. We test our predictive models on the resulting exhaustive same-division pairs for holdout firms (ESDP Holdout) data set that has 61,726 within-industry pairs of firms to understand the role of industrial division in the performance of such models. We note that the predictive models used for testing are trained using the ADP data set after removing all pairs of firms that contain one of the holdout firms.

## 5. Web Metrics of Online Isomorphism

We utilize Web metrics that quantify the online isomorphism between a pair of firms using the overlap in the Web footprints of the firms. The first three metrics measure the online isomorphism using linkage structure and text of websites associated with a pair of firms. Deriving from previous works, we also measure online isomorphism using news stories and general Web pages and use them as control metrics for baseline comparisons. Next, we describe the computation of the Web metrics.

### 5.1. In-Link Similarity

Actors within a social network or, more generally, nodes in a graph have been compared based on the overlap among their neighbors. For example, the structural equivalence metric in social network analysis measures similarity between a pair of actors based on overlap among other actors (e.g., friends, partners) connected to the given pair (Scott 1991). Analogous to a social system, websites of firms make links to other websites (out-links) and have other websites linking to them (in-links). Similarities between competing firms ensued through isomorphism can be expected to lead to firms being linked to by common sites (e.g., a forum discussing the two firms) and the firms linking to common sites (e.g., a trade association or regulatory body in their line of business). Hence, we use overlap between in-links or out-links of a pair of firms as a measure of their online isomorphism, which may reveal their competition.

We note that in-links and out-links present different types of Web footprints. Out-links from a firm's website represent the firm's *own* perception (and disclosure) of its relationships. In contrast, in-links to a firm's website represent perceptions of third parties with respect to their relationships to the firm. In-links, in that sense, provide a possibility of capturing serendipitous relationships that may not be known to (or not revealed by) the firms of interest (e.g., a consumer blog that makes links to two competing firms). The value of in-links in revealing competition has been highlighted previously in the context of Web crawling for business intelligence (Pant and Menczer 2003). We also note that the concept of the in-link similarity metric is similar to the notion of cocitation frequency used in bibliometrics to measure the relationship between documents (Small 1973).

We would like to measure the pairwise overlap in the in-links to websites of firms of interest. Hence, for each of the firms in our test bed, we obtain a list of the top 500 in-link URLs using the Yahoo Boss API.[9] The API provides programmatic access to the Yahoo search engine data. An in-link URL is the URL of a page that contains a hyperlink to the home page on the website (firm) of interest. We omit in-link URLs that are from the same site as the website of interest (i.e., self references). We collect more than a million URLs of pages that link to websites corresponding to the test bed firms.

[9] See http://developer.yahoo.com/search/boss/.

Given a pair of firms (e.g., AMD and Intel in Figure 2(a)), we now have a list of URLs from different websites that have links to the two firms' websites (e.g., pages on microsoft.com and apple.com contain links to home pages of amd.com and intel.com). Because of isomorphism, we expect third-party websites to discuss competitors in similar contexts. For example, a website providing a forum for end users of computer processors that links to AMD's website is indicative of demand-side connections. Hence, if the same Web forum site also links to Intel's website, it may be indicative of the two firms' demand-side substitutability and hence competition. However, it is important to gauge the strength and relevance of these connections. For example, Wikipedia has links to both AMD's and Intel's websites. Is that indicative of their isomorphism and competitive relationship? Wikipedia links to a majority of firms in our test bed, and hence its links are probably less discriminative and hence less relevant in terms of capturing online isomorphism. In contrast, linuxinsider.com links to both AMD's and Intel's sites, and it does not link to many other firms' websites. However, this website has only a single URL/link (among the in-link URLs collected) to AMD's and Intel's sites. Although these lone links may be relevant to the competitive relationship, their strength (due to a low number of links) is very weak.

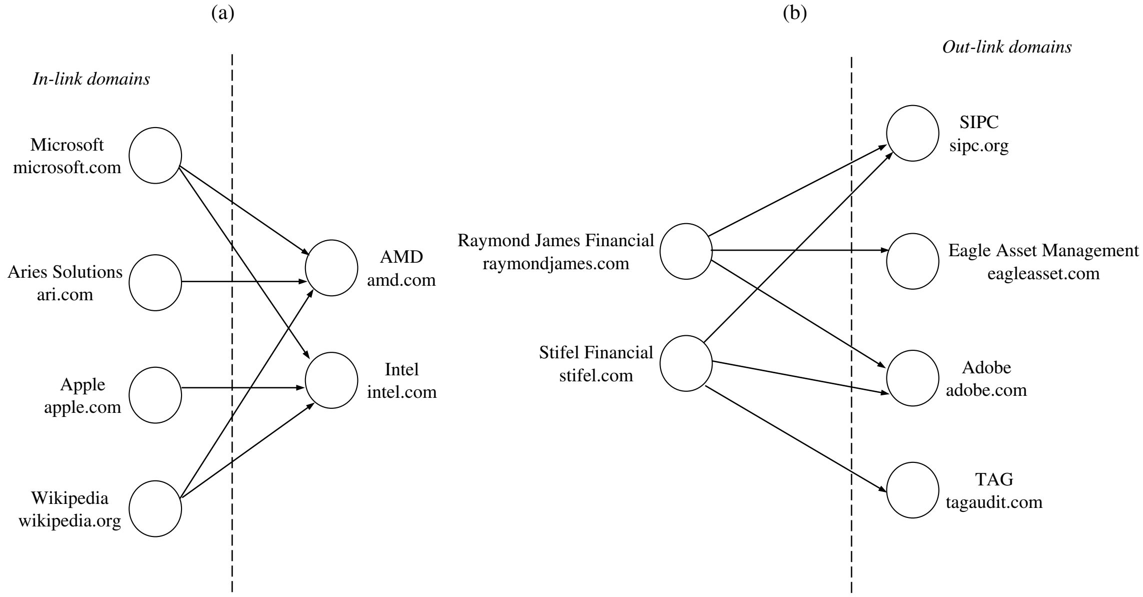
To quantify the concepts described above, we first extract the in-link domains, which are the second-level domains (e.g., microsoft.com for http://www.microsoft.com/) for each of the in-link URLs. We then count the number of times an in-link domain appears among the in-link URLs for a given website (firm). We refer to this count as domain frequency, or DF. For example, 10 URLs from wikipedia.org point to amd.com, and hence the DF of wikipedia.org for AMD is 10. DF is intended to measure the strength of connection between an in-link domain and a firm's website. We further count the number of firms for which a given in-link domain appears (one or more times) among the firms' in-link URLs. We refer to this count as firm frequency, or FF. For example, wikipedia.org appears as an in-link domain for 1,658 firms in our data set, and hence its FF is 1,658. Because of the high FF of Wikipedia, the fact that two firms' home pages are linked from Wikipedia pages is a weak indicator, at best, of their isomorphism. Note that the FF of an in-link domain is the same across firms, whereas the DF changes with the firm (website) of interest. Online Appendix C shows the top 10 in-link domains based on their FF.

It is clear that some domains such as forbes.com link to a large number of firms (websites). Hence, in-links from high-FF domains have lower relevance in terms of indicating isomorphism. In any computation of in-link similarity of two firms, we would need to weigh the in-link domains based on their FF values. Such a weighting function of FF would need to be an inverse of FF, where a higher FF value leads to a lower weight for the corresponding in-link domain. We suggest the following formulation, which we call inverse firm frequency, or IFF, for the weight function, which is analogous to the popular inverse document frequency (IDF) measure in information retrieval:

$$\mathrm{IFF} = \ln \frac{N_c}{FF}, \tag{1}$$

where $N_c$ (= 2,907) is the total number of firms over which the FF of an in-link domain is computed. IFF

**Figure 2** Examples of Overlapping (a) In-Link Domains and (b) Out-Link Domains (Not All Links Are Shown)



increases with decreasing values of FF. IFF can also be written as $-\ln(FF/N_c)$, where $FF/N_c$ can be seen as the probability of an in-link domain having a link to a firm. As the probability of an in-link domain appearing among the in-links of firms increases, it becomes less discriminating, and hence its IFF decreases. Equation (1) is an information theoretic measure akin to Shannon's (1948) entropy measure (Robertson 2004). A similar measure called IDF in information retrieval literature represents the specificity of a word in a collection of documents (Robertson 2004). In our case, IFF represents the specificity of relationship indicated by an in-link domain. In other words, IFF allows us to distinguish a domain that links to most firms, from a domain that links to only a few specific firms. We note that a similar information theoretic measure is used by Hill and Provost (2003) to distinguish between citations made by authors to general references versus more specific references. Although entropy-based measures have been used in different contexts, IFF represents a new application designed specifically for measuring online isomorphism based on firms' website linkage structure.

Each firm $c$ is represented as a vector $\mathbf{w}_c = [w_{c1}, w_{c2}, \ldots, w_{cn}]$ of in-link domain weights, where $w_{cj}$ is the weight for an in-link domain $j$ for firm $c$, and $n$ is the total number of in-link domains across firms. The weight $w_{cj}$ for domain $j$ is computed as a product of DF and IFF

$$w_{cj} = DF \cdot \ln \frac{N_c}{FF}. \tag{2}$$

Again, Equation (2) is analogous to the Term Frequency (TF)-IDF term weight formulation in information retrieval (Manning et al. 2008). Once we have the vector representation for each firm, we find the in-link similarity between two firms $a$ and $b$ as the cosine of the angle (or cosine similarity) between the two corresponding vectors as follows:

$$\mathrm{sim}(a, b) = \frac{\mathbf{w}_a \cdot \mathbf{w}_b}{\|\mathbf{w}_a\| \cdot \|\mathbf{w}_b\|}. \tag{3}$$

The above cosine similarity measures the normalized overlap between in-link domains of two firms while accounting for differing weights of each domain. We note that the in-links that are gathered correspond to Web pages, but we reduce them to Web domains since it allows us to measure the overlap between firms in the space of in-linking domains, which is expected to be relatively less sparse than the space of in-linking pages.

### 5.2. Out-Link Similarity

The out-link similarity between two firms is measured in a manner similar to the in-link similarity. However, instead of using links coming into websites of interest, we now use links going out from the sites (or out-links). As noted earlier, in contrast to the in-links, the out-links represent a firm's perception of other websites related to it. Hence, we can expect this information to be authoritative and complement the third-party information embedded in the in-link similarity. For the purpose of gathering out-links, we crawl up to the first 200 pages[10] from the websites of each of the firms in our test bed. We use a multithreaded breadth-first Web crawler (Pant et al. 2004) that starts from the home page of each firm and follows the links outward in a breadth-first manner to download pages. The crawler follows only those links that are from the same domain as the firm's website and hence avoids downloading pages that are external to the firm.

After downloading Web pages from each firm's website, an HTML parser is used to identify all of the out-links appearing in those pages. For a given firm, a list of out-links is created that corresponds to all Web links that lead to pages outside of the firm's website (domain). Figure 2(b) shows an example of two investment firms and some of their out-link domains. Similar to in-link domains, out-link domains are second-level domains corresponding to out-links of firm websites.

We expect some of the overlapping out-link domains between two firms to be indicative of their isomorphism. In Figure 2(b), both of the firms of interest (left-hand side) have a link to Securities Investor Protection Corporation (SIPC's) website. This may indicate that both of the firms provide products for investors since SIPC provides investor protection (potential demand-side substitutability). The issue of strength and the relevance of the connection between firms of interest and their out-link domains is the same as that discussed for in-link similarity computation. For example, both of the firms in Figure 2(b) have a link to adobe.com, but this overlap may not be indicative of their isomorphism. A large number of firms in our test bed have an out-link to some page on adobe.com (Adobe provides a popular document reader). Hence, we can define firm frequency in this context as the number of firms for which a given out-link domain appears (one or more times) among the firms' out-link URLs. For example, the FF for adobe.com is 1,116. Online Appendix C shows the top 10 out-link domains by firm frequency computed using the out-links. Again, the need for appropriately discounting the high FF domains is clear. We use the IFF function defined in Equation (1) (but now based on FF values computed using out-links) for weighing different domains.

As in the case of in-link similarity, we represent each firm using a vector of weights, where each weight

[10] The intent is to cover pages that are within a few links of the homepage.

corresponds to an out-link domain, and the weight is computed using Equation (2). (Note that DF and FF are computed using the out-links.) Finally, the out-link similarity between two firms is measured as the cosine similarity between the out-link weight vectors of the two firms.

### 5.3. Text Similarity

The website of a firm describes various aspects of its business. The text in the first few pages on a firm's website can be considered a type of self-description by the firm. As explained in the previous section, we crawled up to the first 200 pages from each firm's website. We now concatenate the text from these crawled pages and use it as a self-description of the firm. Based on the discussion in §3, we may expect that competing firms will describe themselves similarly. To measure the similarity between self-descriptions of two firms, we use one of the standard TF-IDF representations (Manning et al. 2008) from information retrieval. The words in each of the self descriptions are identified. We remove stop words or common words such as "and," "or," "the," etc., and also ignore numeric or alphanumeric words (e.g., "67.2," "a12"). We compute the weights of each of the remaining words as follows:

$$t_{kc} = \left(0.5 + \frac{0.5 \cdot f_{kc}}{\max_{k' \in T_c} f_{k'c}}\right) \cdot \ln \frac{|E|}{d_k}, \qquad (4)$$

where $t_{kc}$ is the weight of the word $k$ in self-description of firm $c$, $f_{kc}$ is the frequency of the word $k$ in the self-description of the firm $c$, $T_c$ is the set of words appearing in the self-description of the firm $c$, $d_k$ is the document frequency of the word $k$, and $|E|$ is the set of pages over which document frequencies are computed. The document frequency of a word is the number of pages in which the word appears. We compute document frequencies over all of the pages crawled across all of the firms in our test bed. A word that tends to appear in a large number of pages can be considered to be less specific and hence probably less meaningful while measuring the similarity between self-descriptions of two firms. Equation (4) is one of the standard TF-IDF formulations (Manning et al. 2008). Each of the self-descriptions is represented as a vector of word weights, and their similarity is measured through cosine similarity as depicted in Equation (3).

### 5.4. News Count

If two firms are mentioned in the same piece of news, it may indicate some type of connection between the two. This has been a fundamental assumption behind some of the previous works that have utilized such co-occurrence to suggest Web metrics that may indicate interfirm relationships (Bernstein et al. 2002; Ma et al. 2009, 2011). We now use a metric that utilizes the fundamental assumption in these previous works that co-occurrence in news stories indicates a potential interfirm relationship. In particular, using the Yahoo Boss API, we obtain the number of news stories (in a month) in which two firms' names co-occur. We call this measure *name news count*. For this purpose, we first canonicalize the firm names by removing words such as "Inc." and "Corp.," and then concatenate the two names to create a keyword that is searched against news stories using the Yahoo Boss API. The API returns the number of news pages that match the keyword (this number is similar to the count of results that typically appears at the top of search engine result pages). Thus, we obtain the number of news stories with the names of the two firms. As an additional measure, we also count the number of news stories that contain the tickers of the two firms (instead of names). We call this measure *ticker news count*.

### 5.5. Search Engine Count

Bao et al. (2008) utilize the co-occurrence of firm names among search engine results as a fundamental step in their process of mining competitor relationships. Also, co-occurrence of words in search results has been used previously to successfully discover synonyms for automatically answering questions from standardized English tests such as the TOEFL (Turney 2001). In essence, co-occurrence of words can often indicate semantic similarities between the words (and the concepts that they represent). In our case, the words of interest represent firms, and hence we expect to uncover similarities or online isomorphism between firms based on the co-occurrence of their names or tickers in search results. Hence, in a manner similar to news count, we query Yahoo Boss API's Web search service with firm names and tickers of the two firms of interest and obtain *name SE count* and *ticker SE count* as metrics that we utilize in our analysis. Again, the API returns the data on the number of Web pages that match the keyword containing the firm names (or tickers).

This is the first work to suggest a variety of Web metrics for measuring online isomorphism and exploiting it for the problem of competitor identification. However, we note that metrics based on news and search engines have been individually utilized in the past for exploring interfirm relationships. Hence, we will use the news count and search engine count as control metrics.

## 6. Empirical Support for Online Isomorphism

While suggesting the aforementioned web-based metrics, we hypothesize that they capture online isomorphism. In other words, we expect these metrics to behave differently for firms that are competitors versus those that are not. We now statistically test this hypothesis for each of the metrics of online isomorphism that we described earlier. We note that this is the first direct test of online isomorphism that is an online equivalent of the sociological phenomenon of isomorphism between competing firms (Hannan and Freeman 1977, DiMaggio and Powell 1983).

We test the online isomorphism hypothesis through a set of hypotheses corresponding to each metric. More specifically, we test the hypotheses that the mean of each of the Web metrics of online isomorphism is higher for competitors than for noncompetitors (these correspond to H1, H2, H3, H4A, H4B, H5A, and H5B in Table 2). To test these hypotheses, we first use the ADP data set. The Web metrics are computed for the 16,485 competitors (pairs of firms) and 16,485 noncompetitors (pairs of firms) in the data set. Table 2 shows the results of the various two-tailed *t*-tests for significant difference between means of Web metrics for competitor and noncompetitor pairs. The table also shows the average values of the various Web metrics for competitors and noncompetitors along with the respective standard errors. All of the hypotheses other than H4B and H5B are supported at the 0.01 significance level. H5B is supported at the 0.05 level, whereas H4B can be rejected. In other words, the three metrics suggested by us, on average, behave differently for competitors versus noncompetitors. This is also true for the news and search engine count metrics (using firm names) that are derived from fundamental assumptions in previous literature. In other words, online isomorphism is indeed present, and it can be captured through Web metrics.

To understand the role of industrial division in online isomorphism, we test the hypotheses corresponding to each of the Web metrics on the SDP data set where the two firms in each (competitor and noncompetitor) pair are from the same industrial division. The Web metrics are computed for the 3,561 competitors (pairs of firms) and 3,561 noncompetitors (pairs of firms) in the data set. Similar to Table 2, Table 8 in Online Appendix E shows the results of the various two-tailed *t*-tests for significant difference between means of Web metrics for competitor and noncompetitor pairs using the SDP data set. All of the hypotheses other than H4B and H5B are again supported at the 0.01 significance level. In other words, the signals of online isomorphism are strong even when we control for the industrial division by making sure that firms in a pair are from the same division.

We note that H4A and H5A are significantly supported in both ADP and SDP data sets (at the 0.01 significance level), whereas H4B can be rejected for both ADP and SDP data sets. Furthermore, H5B can be rejected for the SDP data set and is only weakly supported in the ADP data set. In other words, ticker-based news and search engine counts are too noisy to provide a reliable signal of online isomorphism. On the other hand, similar counts based on firm names provide strong signals of online isomorphism. Hence, we will use just the firm name-based news and search engine counts for further analysis and leave out the ticker-based counts. Online Appendix D provides the pairwise correlation, based on the ADP data set, between the five Web metrics identified for further analysis. We find that almost all of the correlations are weak ($< 0.2$) other than a moderate correlation (0.53) between news count and SE count. Hence, the Web metrics suggested by us potentially capture different aspects of online isomorphism between firms. In other words, we can expect these metrics to complement each other for the problem of competitor identification.

**Table 2 Results of Hypothesis Testing for Various Web Metrics Along with the Average Values of the Metrics for Competitors and Noncompetitors (ADP Data Set)**

| Hypothesis | Class | Mean | Std. error | Sig. (2-tailed) |
|---|---|---|---|---|
| H1: In-link similarity | Competitor | $4.93 \times 10^{-2}$ | $8.60 \times 10^{-4}$ | 0.000 |
| | Noncompetitor | $7.55 \times 10^{-3}$ | $2.78 \times 10^{-4}$ | |
| H2: Out-link similarity | Competitor | $2.05 \times 10^{-2}$ | $7.23 \times 10^{-4}$ | 0.000 |
| | Noncompetitor | $1.31 \times 10^{-2}$ | $5.75 \times 10^{-4}$ | |
| H3: Text similarity | Competitor | $6.13 \times 10^{-2}$ | $5.20 \times 10^{-4}$ | 0.000 |
| | Noncompetitor | $2.61 \times 10^{-2}$ | $3.57 \times 10^{-4}$ | |
| H4A: Name news count | Competitor | $2.66 \times 10^{1}$ | 1.81 | 0.000 |
| | Noncompetitor | 3.96 | $5.64 \times 10^{-1}$ | |
| H4B: Ticker news count | Competitor | $2.54 \times 10^{2}$ | $1.15 \times 10^{2}$ | 0.121 |
| | Noncompetitor | $7.27 \times 10^{1}$ | $2.27 \times 10^{1}$ | |
| H5A: Name SE count | Competitor | $7.78 \times 10^{4}$ | $4.76 \times 10^{3}$ | 0.000 |
| | Noncompetitor | $1.79 \times 10^{4}$ | $1.57 \times 10^{3}$ | |
| H5B: Ticker SE count | Competitor | $2.19 \times 10^{5}$ | $2.35 \times 10^{4}$ | 0.041 |
| | Noncompetitor | $1.47 \times 10^{5}$ | $2.65 \times 10^{4}$ | |

*Note.* Sig., Significance.

## 7. Predictive Models

Our empirical analysis suggests that the five Web metrics of online isomorphism identified earlier may act as good predictors for competitor identification. Using the metrics computed for competitor and noncompetitor pairs (firms), we build predictive models where inputs to a model are the online metrics (we later use two offline metrics for comparison) and the output is a class labeled C (competitor) or NC (noncompetitor). We use two different predictive modeling techniques: the C4.5 decision tree and logistic regression. These modeling techniques are popular in the data-mining and econometric modeling literature. Using the ADP data set, we train the models using 66% of randomly selected data (pairs) and hold out the remaining data for testing. To maintain the disjoint nature of the training and testing data sets, we remove the pairs of competitors from the testing data whose reverse instances[11] appear in the training data. The testing data represent the unknown (competitor/noncompetitor) information for the models.

### 7.1. Skewed Data Sets

To understand the sensitivity and relative performance of models on data with different levels of skewness, starting with the ADP data set, we create data sets with different ratios of competitors and noncompetitors. The data sets with different ratios represent different levels of information deficit. A data set with a 1:1 ratio of competitors and noncompetitors is a balanced data set where it is equally likely to find a competitor and a noncompetitor (i.e., 50% prior probability of each class). If the managers or analysts are not well informed about the external and internal environments of the focal firm or the candidate competitors, they will need to consider a much larger set of target firms than the true competitors. In such scenarios, the data are much more skewed toward noncompetitors.

To create the skewed data sets, we first split the data into training and testing as described earlier. Then we randomly filter out data instances to obtain a desired skewed competitor to noncompetitor ratio (1:1, 1:2, 1:5, and 1:10) within the training and testing data sets. The training and testing data sets have the same ratio (i.e., class distribution). A ratio of 1:5 indicates that for every competitor (pair) in our training and testing data sets, we have five noncompetitors (pairs).

### 7.2. Measuring Predictive Performance

We repeat the training and testing process 50 times using each of the models. In other words, we randomly divide the overall data into training and testing data sets 50 times, and each time we train the models using the training data and observe their performance on the testing data. This cross-validation mechanism is called repeated random subsampling validation. Hence, we obtain 50 observations on the performance of each modeling technique for each data set (skewness level). Our results are based on average testing data performance from the 50 repeated experiments for each of the modeling techniques. We use the Weka machine learning software (Witten and Frank 2005) for our experiments.

The performance of a modeling technique is measured using the following standard measures:

1. *Mean precision*. The precision of a given technique is the fraction of firm pairs identified as competitors by the technique that are actual competitors. The precision is computed using the testing data for each test/train run, and the mean precision is computed over the 50 runs.[12]

2. *Average recall*. The recall of a given technique is the fraction of actual competitor pairs that are correctly identified as competitors by the technique. The recall is computed using the testing data for each test/train run, and the average recall is computed over the 50 runs.

3. *Average F-measure*. For a given technique, precision and recall measures often present a trade-off (i.e., attempts to increase precision can lead to lower recall and vice versa). Therefore, the $F$-measure, a harmonic mean of precision and recall, is popularly used to integrate precision and recall into one measure as follows:

$$F = \frac{2 \cdot P \cdot R}{P + R}, \tag{5}$$

where $P$ is precision and $R$ is recall for the given technique. The $F$-measure is computed using the testing data for each test/train run, and the average $F$-measure is computed over the 50 runs.

4. *Average area under the curve*. The area under the curve (AUC) represents the area under the receiver operating characteristic (ROC) curve. The ROC curve in the current context plots the fraction of actual competitor pairs that are correctly identified as competitors (i.e., recall) against the fraction of noncompetitor pairs that are incorrectly classified as competitors. The two values may vary as the decision threshold of the classifier is varied, thus providing the ROC curve. The greater the AUC, the better the performance of a predictive model under various thresholds. The AUC is

[11] "Exxon is a competitor of Chevron" is the reverse instance of "Chevron is a competitor of Exxon." This is not always true since competitor relationships can be asymmetric.

[12] We use the term *mean precision* to avoid confusion with another average precision measure in the literature (Manning et al. 2008).

measured using the testing data for each test/train run, and the average AUC is computed over the 50 runs.

5. *Average accuracy*. The accuracy of a given technique is the fraction of firm pairs that are correctly classified as competitors or noncompetitors. The accuracy is computed using the testing data for each test/train run, and the average accuracy is computed over the 50 runs.

We note that while precision and recall present a trade-off, the *F*-measure, AUC, and accuracy provide aggregate performance measures. Also, when the class distribution in data sets is skewed, the *F*-measure and AUC (Weiss and Provost 2003) provide more information than accuracy. Hence, we consider the *F*-measure and AUC as the more important performance indicators for the purposes of this work. However, we will augment them with the other three measures to provide a richer set of performance measurements.

We tease out the predictive power provided by the three Web metrics suggested by us over the two control metrics that have been suggested previously in the literature. To rigorously achieve this comparison, we build predictive models using (1) search engine count, (2) news count, (3) news and search engine counts, (4) three new metrics based on content and links, and (5) all five online metrics. We present results with both logistic regression and the C4.5 decision tree technique. We note that there are no counterparts in the literature to models using (3)–(5) described above.

## 8. Performance

### 8.1. Online Metrics

Figure 3 shows the performance of the five decision-tree-based models for different competitor to noncompetitor ratio data sets. It is very clear from the different measures of predictive performance that the most effective strategy is to use a model based on all five online metrics discussed earlier. However, it is also important to note that the second most dominant strategy is to use the three new metrics proposed by us, which provide a clear predictive performance advantage over the control metrics derived from the literature. In other words, not only do the new metrics provide additional predictive value when combined with the control metrics, they largely provide better predictive performance than the control variables. In particular, using all five metrics provides an 11% to 1,704% better average *F*-measure, a 10% to 32% better average AUC, and a 1% to 13% better average accuracy than using the news and SE counts (together) across the data sets with different levels of skewness. Using the three new online metrics provides a 4% to 1,449% better average *F*-measure, 7% to 26% better average AUC, and 1% to 9% better average accuracy than using the news and SE counts (together) across the data sets with different levels of skewness. All of these improvements achieved by using all online metrics or just the three new metrics over using the news and SE counts (together) are statistically significant ($t$-test, $p < 0.001$).

Figure 8 in Online Appendix F shows the performance of the five logistic regression-based models for different competitor to noncompetitor ratio data sets. Again, using all five online metrics together is the dominant strategy, whereas using the three new metrics is a very close second-best strategy. We also note that the performance of the best decision-tree-based model (using all online metrics) is statistically significantly better than the best logistic regression-based model (using all online metrics) on average *F*-measure, average AUC, and average accuracy across data sets with different levels of skewness (the only exception is observed for the 1:10 data set for which logistic regression has a higher average AUC). Given the observed advantage of using decision trees for this problem and with the intent of reducing redundant experiments, we focus on primarily using the decision tree models for the rest of this paper.

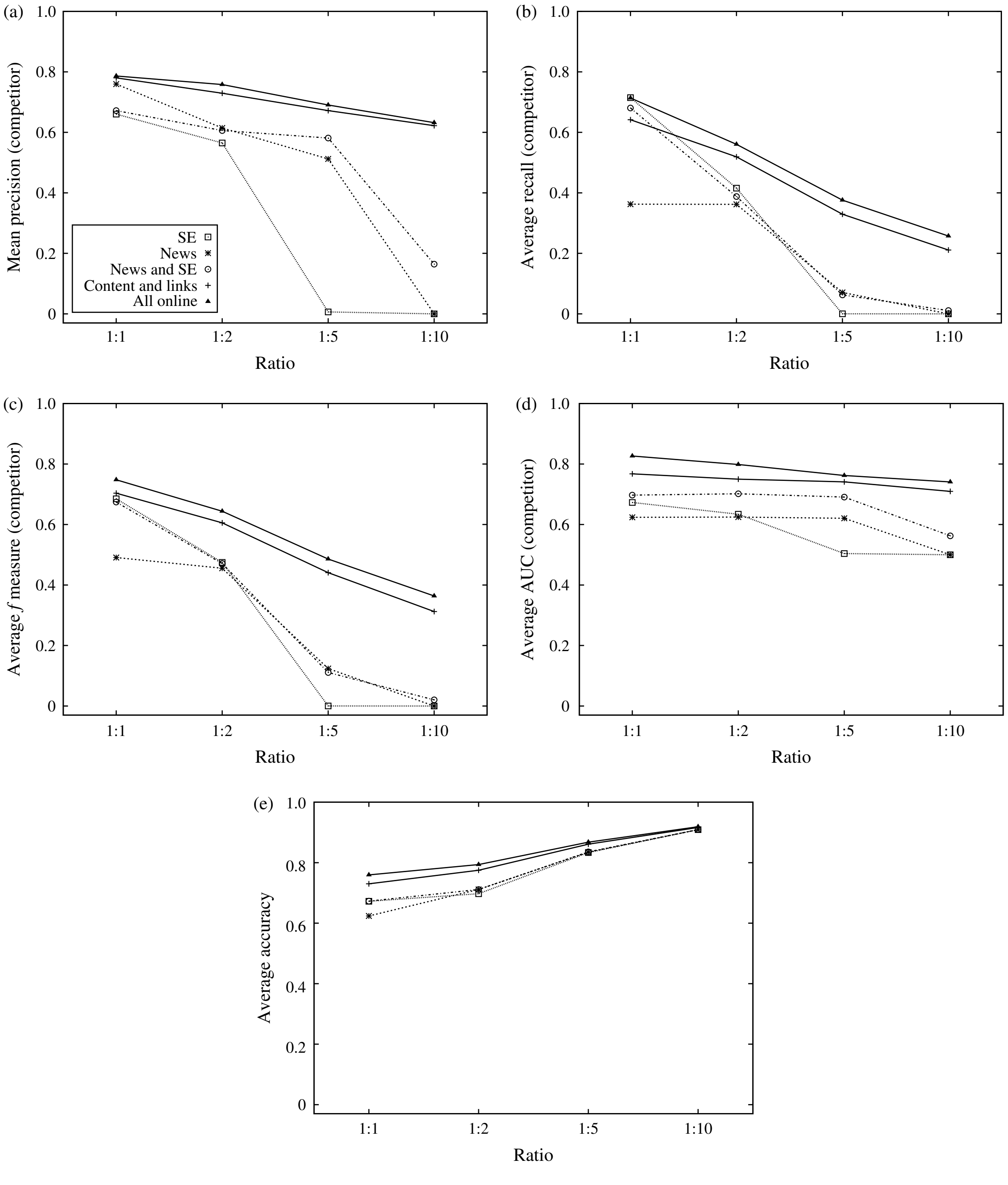
To shed some light on the nature of decision tree models with five online metrics, we provide visualizations of a couple of such trees. Figure 9 in Online Appendix G shows two decision trees based on a random (training/testing) split using a 1:1 ratio data set and a 1:10 ratio data set. We note that the decision trees shown there are simplified (i.e., made more readable) by adding a constraint on the minimum number of instances in the leaf nodes while building the models.

To summarize, we find that using all of the suggested Web metrics provides a clear advantage over using only the control metrics for all of the data sets with different skewness levels (ratios of 1:1, 1:2, 1:5, and 1:10). Moreover, using the three new metrics alone provides a better predictive performance than the control metrics. These results tease out the strong predictive value of the new metrics of online isomorphism suggested by us. The results also provide a rigorous argument for an integrated use of complementary metrics of online isomorphism in models that attempt to identify competitor relationships. Both of these aspects (new metrics and an integrated framework) serve as important contributions to the related literature.

### 8.2. Benchmarking with Offline Metrics

While comparing our newly suggested online metrics against the control metrics (i.e., news count and

**Figure 3** Performance of Decision Tree Models Using All Five Web-Based Metrics (All Online), Only the Three New Metrics (Content and Links), Only the News Count (News), Only the Search Engine Count (SE), and the Search Engine Count with the News Count (News and SE)



SE count) derived from previous literature, we keep the modeling methods the same. This is an important design choice that allows us to make fair and consistent comparisons between the metrics derived from different information sources. By keeping the modeling methods (decision tree or logistic regression) constant, we are able to tease out the predictive improvements due to the newly suggested online metrics and quantify their relative performance compared to the control metrics. We now extend the control metrics to include metrics based on two popular offline characteristics of firms—SIC codes and market cap.[13] We obtain the offline data from the Center for Research in Security Prices (CRSP). Similar to the online metrics, the offline metrics are associated with a pair of firms. The first offline metric is *SIC similarity*, and it is a binary value that has a value

[13] Market cap is the share price times the number of shares outstanding.

of 1 when the SIC codes of a pair of firms are the same and 0 otherwise. The second metric is *market cap difference*, which is simply the difference between the market cap of the focal firm and the target firm when considering a pair of firms. The market cap difference can be positive or negative. We use these two offline metrics as additional predictors.

We may expect that if the SIC similarity of a pair of firms is 1, the firms are likely to be competitors. On the other hand, the role of market cap difference in determining competitor relationship may be more complex. It is possible that firms of similar market cap tend to be competitors; however, large firms may pose a competitive threat to a wide variety of smaller firms. Nevertheless, it is important to benchmark our approach of using online metrics against these offline control metrics given their usefulness and popularity in characterizing public firms. We note that the limited literature on automated competitor identification does not provide direct predictive performance comparisons against these (or other) offline metrics. Ma et al. (2011) consider industry sectors as defined by Yahoo Finance! as an additional feature in their models based on online news data but do not compare directly with an offline metric. We feel that this lack of comparison is an important gap in the literature since online (more generally, digital) space does not exist in isolation from the offline (or social) realm. Hence, comparisons against offline metrics provide a necessary and practically useful benchmark.

Figure 4 shows the prediction performance of decision-tree-based models using both of the offline metrics (offline), just SIC similarity (SIC), all online metrics (all online), and all online and offline metrics together (all online and offline). We find that the models using just SIC similarity can achieve the highest precision among all of the evaluated models. However, the same SIC-based models also produce the lowest recall. In other words, if the SIC codes of two firms are the same, the firms are highly likely to be competitors (causing high precision). At the same time, many competitors do not share the same SIC code (causing low recall). If we focus on the aggregate performance measures (i.e., Figures 4(c)–4(e)), we find that the dominant strategy is to use both online and offline metrics. The second-best strategy is to use all of the online metrics. The offline metrics provide a strong benchmark, but we find that online metrics are competitive with or better than this benchmark. It is also interesting to note that combining the online and offline metrics outperforms their individual usage. In other words, offline and online signals of competitor relationships are complementary in nature. We would like to also highlight that online signals are more generally applicable, since firm-level variables such as SIC codes and market caps are only available for public firms.

In Online Appendix H we consider a nuanced aspect of competitor relationships, i.e., the symmetry of such relationships. When firm A is seen as a competitor of firm B, and firm B is also recognized to be a competitor of firm A, the competitor relationship between firms A and B can be called symmetric. However, if firm A is seen as a competitor of firm B, but firm B is not recognized to be a competitor of firm A, the competitor relationship between the two firms is asymmetric. We provide results for a predictive model that differentiates between asymmetric and symmetric competition. To the best of our knowledge, this is the first attempt at using predictive models to identify symmetric and asymmetric competition. We find that predictive models with reasonable performance can be built for this nuanced dimension of the competitor relationship.

In addition, in Online Appendix I we tease out the performance of the decision tree model by dividing the testing data set based on industrial division. We note that the performance for different industrial divisions can vary from the overall performance seen in Figure 3. We study the role of industrial division in detail in §9.

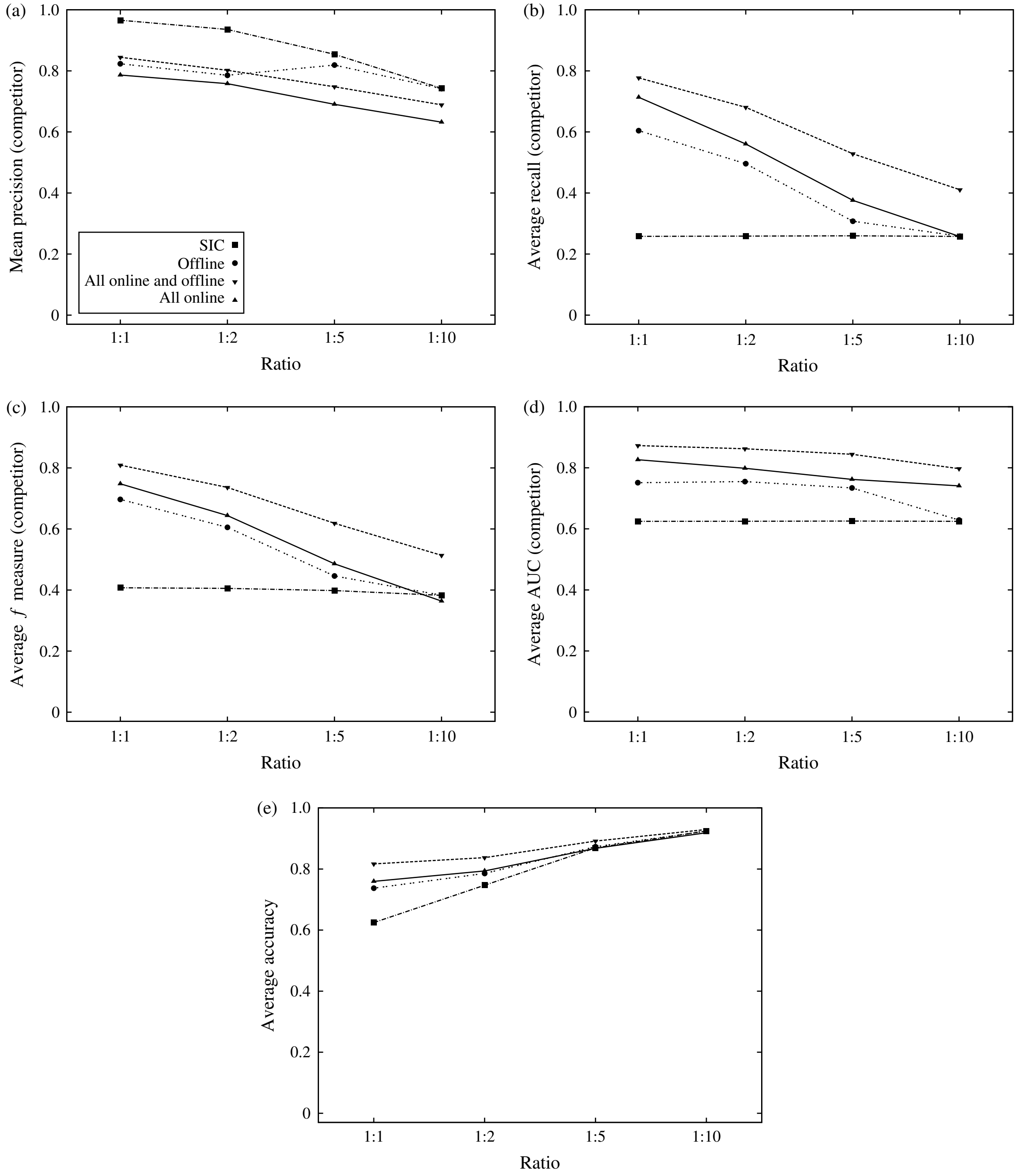
## 9. Role of Industrial Division

The firms in our data set come from various industries, and we now explore the role of these industries in determining the performance of our models. As noted earlier, the U.S. DOL divides firms into 10 divisions (e.g., “Mining,” “Construction,” “Services”) using the SIC codes of the firms. Online Appendix B shows the specification that can be used to map firms to their industrial divisions using their respective SIC codes.

### 9.1. Same-Division Competitors and Noncompetitors

Using 2,673 firms in the ADP data set, we find that, on average, 80% of a firm’s competitors fall into the same industrial division; hence, we ask, *Can the predictive models differentiate between competitor and noncompetitor pairs of firms if the two firms are from the same industrial division?* Firms within an industrial division may show some similarities due to shared industrial space without being competitors. It is hence possible that differentiating between competitors and noncompetitors would be harder when the two firms being considered are from the same industrial division. On the other hand, by considering pairs of firms within an industrial division, we decrease the search space to a region where there is a greater likelihood of finding competitors. These competing arguments may suggest that it is not clear which of these

**Figure 4** Performance of Decision Tree Models Using the Offline and Online Metrics



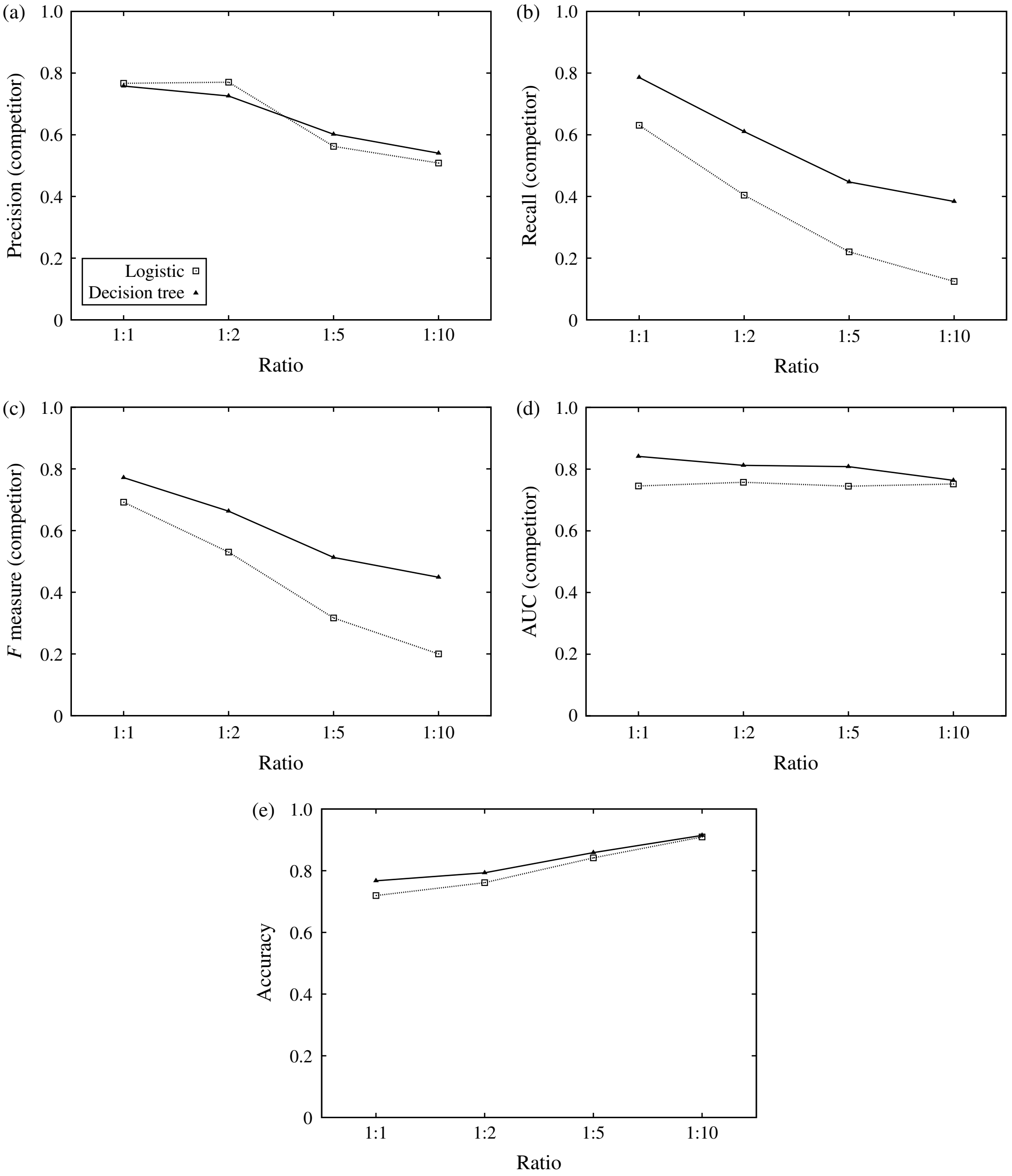
two factors will play a bigger role in negatively or positively affecting the predictive performance. However, our empirical analysis with the SDP data set (see Online Appendix E) suggests that even if we consider only pairs of same-division firms, on average there is a significant difference between the five Web metrics of online isomorphism computed for competitors as opposed to noncompetitors. In other words, we can expect the predictive models of competitor identification suggested earlier to be effective even when the competitors and noncompetitors are restricted to same-division firms.

To explore the question of predictive performance for same-division firms directly, we use the ESDP Holdout data set described in §4. The ESDP Holdout data set has 100 randomly selected holdout firms from our original data set of 2,673 firms. For each of the holdout firms, we identify an exhaustive list of all firms in our data set that fall into the same industrial division (as the holdout firm). We then prepare pairs of firms by taking each holdout firm and matching it with a same-division firm. This provides us with 61,726 pairs of firms. Next we obtain and compute the five Web metrics for each pair of firms. Using the process described in §7.1, we prepare data sets with different proportions of competitor and noncompetitor pairs. However, this time each of the competitor and noncompetitor pairs contains firms from the same industrial division.

We note that the holdout firms represent firms for which no a priori information is available for building a classification model. More specifically, we take all of the data that are available for competitor and noncompetitor pairs (as described in §4) in the ADP data set and remove any instances that contain the holdout firms as a focal or target firm. We use these filtered data (with the five Web metrics and the class labels) for training a C4.5 decision tree model and a logistic regression model. We test these predictive models on the data sets with various skewness levels (i.e., competitor to noncompetitor ratios) that are derived from the ESDP Holdout data set.

Figure 5 shows the performance of decision tree and logistic regression models for data sets with different levels of skewness. We first observe that the decision tree models largely outperform the corresponding logistic regression models. More importantly, we find the performance of the decision tree models in classifying same-division competitor and noncompetitor pairs (Figure 5) to be comparable to

**Figure 5** Same-Division Competitors and Noncompetitors: Performance of Decision Tree and Logistic Models Using the Online Metrics



the performance of similar decision tree models in classifying more general competitor and noncompetitor pairs (Figure 3). Overall, the two factors that we alluded to earlier seem to either cancel out or positively affect the predictive performance. In summary, the predictive models that we propose are also adept at classifying same-division competitor and noncompetitor pairs.

### 9.2. The Ranking Problem

A binary classification model (i.e., competitor/noncompetitor) is a natural setup for the automated competitor identification problem. However, in certain situations it may be necessary to rank firms by their likelihood of being a competitor to a given focal firm. For example, when an analyst is just beginning to explore a given firm's competitive environment, she may want to consider all of the firms in a firm's industrial division. An effective tool that can rank the firm by their likelihood of being a competitor would be very useful for the analyst to prioritize an otherwise exhaustive search over a large number of firms. We now explore our decision tree classification model in such a rank-order setting and measure its effectiveness. The gold standard data (obtained using the Hoovers API) do not rank competitors. However, we can measure the effectiveness of the ranking provided by the suggested predictive models through measures such as *lift* as described later. The exploration of the ranking problem adds further to the robustness of the current study.

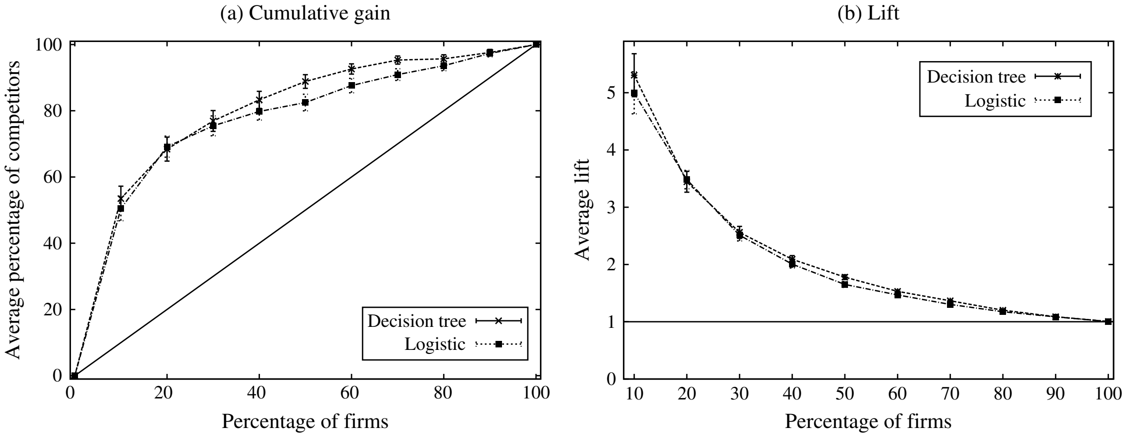
As noted earlier, we find that, on average, 80% of a firm's competitors fall into the same industrial division. Hence, although it may not cover all competitors, the industrial division is a good place to start looking for competitors for an analyst who is beginning to explore a given firm's competitive environment. However, only a small percentage ($< 2\%$ on average) of the same-division firms are competitors of a focal firm of interest. In other words, exploring firms within the industrial division of the focal firm for competitor identification is a "needles in a haystack" problem. Any tool that can increase the number of "needles" found while covering a certain amount of "haystack" would be valuable. The search over the firms in the same industrial division can be made more effective by prioritizing the exploration through meaningful ranking of the firms.

The setup for studying the ranking problem uses the ESDP Holdout data set. Similar to §9.1, the C4.5 decision tree model is trained with the filtered ADP data set (with the five online metrics and class labels) where the instances with holdout firms have been removed. For a given holdout firm, we then use the trained decision tree model to classify each pair (with the holdout firm as the focal firm) in the ESDP Holdout data set as a competitor or noncompetitor. We obtain the probability distribution that the model provides over the two classes (i.e., competitor/noncompetitor) and use the probability of being in the competitor class as a score for ranking all same-division firms. Hence, the probability distribution can allow us to rank each of the firms that are in same industrial division as the holdout firm by the likelihood of the firm being a competitor of the holdout firm. We note that the decision tree model here is similar to that used in §7. However, we now use the probability of a pair of firms falling in the competitor class (as opposed to absolute class membership) for ranking. The use of output probabilities of classifiers for ranking is popular in direct marketing and customer relationship management applications (Padmanabhan et al. 2006).

We need a measurement to evaluate the effectiveness of the ranking achieved through the probabilities generated by the decision tree model. We note that the problem of ranking firms in the current setting is not very different from ranking potential customers in a direct marketing campaign. A popular evaluation measure in the latter context is lift (Hughes 2000, Padmanabhan et al. 2006), which provides valuable insight into the return on investment of the model used for ranking. In particular, the lift measures the effectiveness of a given ranking scheme vis-á-vis a random exploration of the candidates. Hence, it can provide a good understanding of time and/or cost savings through a ranking scheme after covering a certain percentage of the candidates. In our case, the candidates are the same-division firms, and the ranking is provided by the decision tree model.

We draw two different plots to understand the ranking effectiveness of our model: the cumulative gains chart and the lift chart. Cumulative gain measures the percentage of same-division competitors that are identified by the decision tree model. The measurement can be made after covering a given percentage (top 10%, 20%, and so on) of the total same-division firms. Cumulative gain is essentially the recall after covering $N\%$ of the candidate firms through a given ranking scheme. The lift is defined as the fraction of ranked firms that are competitors as a ratio of the baseline rate (i.e., prior probability of same-division competitors). Again, the measurement can be made after covering a given percentage (top 10%, 20%, and so on) of the total same-division firms. We find that six of the 100 holdout firms do not have any competitor within their industrial division; the cumulative gain and lift cannot be defined for these firms, and hence we leave these firms out of our measurements. We average the cumulative gain and lift over the remaining 94 holdout firms.

**Figure 6** **The Average Cumulative Gain and the Average Lift Provided by the Predictive Model That Ranks Same-Division Firms for Each Holdout Firm**



*Note.* The error bars indicate ±1 standard error.

Figure 6(a) shows the cumulative gains curves for the holdout firms. The horizontal axis shows the percentage of same-division firms covered, whereas the vertical axis shows the cumulative gain. The diagonal line represents the baseline rate of finding competitors among the candidates. The higher the cumulative gains curve above this baseline, the better the ranking scheme. To illustrate, for a given holdout firm, after covering the first 20% of the same-division firms, the baseline rate would suggest that the analyst should find 20% of its same-division competitors. However, on average, the ranking provided by the decision tree model helps identify more than 68% of the same-division competitors among the top 20% of the same-division firms ranked by the decision tree. This represents a more than three times higher return on investment (in terms of effort) for a manager using our model compared to the baseline rate. The cost or effort savings are more obvious from the lift chart seen in Figure 6(b). The horizontal line with a lift of 1 represents the baseline rate. We find that the ranking provided by the decision tree model provides a significant advantage allowing the analyst to significantly reduce the search cost by covering a small percentage of firms to explore a much larger number of competitors than that suggested by the prior probability of finding competitors among the same-division firms. For example, as seen in Figure 6(b), covering the first 10% provides a lift of more than five times compared to the prior. Figure 6, (a) and (b), includes the ranking performance based on a logistic regression model as well. The logistic regression model is used in exactly the same manner as the decision tree model. We find that the logistic regression performs marginally worse than decision tree. Online Appendix J provides a more detailed exploration of ranking effectiveness of the decision-tree model based on the division and market cap of firms.

## 10. Predicting New Future Competitors

The results and analyses using the SDP, ADP, and ESDP Holdout data sets show us that the websites of firms and their surrounding linkage structure provide strong cues on the contemporaneous competitive relationships of firms. We now investigate whether the proposed metrics of online isomorphism computed at a certain time can predict future competitors that were unknown (based on the gold standard source, i.e., Hoover’s) at the time that the metrics were computed. For this purpose, we start with the 100 random firms in the ESDP Holdout data set. For these firms, we obtain competitors from Hoover’s five years after the Web metrics were derived.[14] Using this new Hoover’s data, we identify *new future competitors* as those competitors of these 100 firms that were not listed in Hoover’s data set from five years ago. Since our Web metric computations are based on the 2,678 focal firms from the Russell 3000 index, we restrict the list of new competitors to the same set of firms. A total of 196 new future competitors were obtained for the 100 holdout firms from the new Hoover’s data that lead our Web metrics (and hence models) by five years. Whereas we have seen effectiveness of the decision tree model based on the proposed Web metrics in identifying contemporaneous competitors, we now measure how many of the new future competitors (from five years later) the same model will recognize. Like in §9.1, the C4.5 decision tree model is trained with the filtered ADP data set, where the instances with the 100 holdout firms have been removed. We then apply this trained model to check how many of the new future competitors of the 100 holdout firms the model can identify (as competitors to those firms).

[14] The new data on competitors were obtained manually from Hoover’s online portal (http://www.hoovers.com).

**Table 3 Predicting New Future Competitors**

| Predictors | Number/percentage of new future competitors identified (%) |
|---|---|
| SIC | 20/10.2 |
| Offline variables | 75/38.3 |
| Online variables | 97/49.5 |
| Online and offline variables | 89/45.4 |

To benchmark the performance of the five online metrics, we train multiple models using the online and offline variables, similar to §8.2. Table 3 lists the performance of the models in terms of the percentage and the number of new future competitors identified with different variables.

The best performance in identifying the new future competitors is achieved using only the online variables. An example of this performance is provided in Online Appendix A. A focal firm (Harmonic) and a few target firms are shown in Figure 7(b) (Online Appendix A). At the time that the online footprints for the firms in this example were collected, our gold standard data from Hoover's did not identify the firm named Harris (HRS) as a competitor for Harmonic (HLIT). However, our predictive model based on the online isomorphism cues associated a high probability of 0.8 that Harris is a competitor of Harmonic. Interestingly, when we obtained new Hoover's data, five years later, it did include Harris as a competitor of Harmonic.

We also observe that although the offline variables complement online variables in identifying contemporaneous competitors (see Figure 4), they are apparently too noisy to help in finding new future competitors (see Table 3). Also, we know from the ADP data set that, on average, 27% of a firm's competitors have the same SIC code as the firm. However, as shown in Table 3, our model based on just the SIC code correctly identifies only 10% of new future competitors. In other words, new future competitors are less likely to come from the same SIC code compared to the current competitors. One reason for this can be that firms expand or refocus their businesses over time, leading to new competitors that are not evident through their current SIC codes. Also, market values are considerably dynamic and hence may not provide reliable signals much into the future. These may explain why these offline variables do not provide a useful complementary signal to the online variables in identifying new future competitors. On the other hand, firms leave overlapping online footprints in terms of similar content on the firms' sites, similar linkages, and co-occurrences in news and other websites that can provide better leading indicators of competition. This further accentuates the need to consider online isomorphism while investigating competitive relationships.

## 11. Conclusion

Motivated by the sociological notions of isomorphism between competing firms, we explore and provide the first direct evidence of a parallel phenomenon of online isomorphism in the Web footprints of competing firms. We suggest new metrics of online isomorphism based on the content and linkage structure of firms' websites. We then utilize the presence of online isomorphism for the competitor identification problem. Competitor identification has been highlighted as a critical and challenging step in competitive analysis and strategy, but there is limited literature on the automatic identification of competitors. We use online metrics as inputs in predictive models that classify pairs of firms as competitors or noncompetitors. We find the resulting predictive models provide high accuracy, *F*-measures, and AUCs. The models also indicate that using a variety of Web metrics, as suggested by us, provides a clear benefit compared to just using the individual control metrics that are derived from previous literature. The benefit is observed for data sets with different proportions of competitor and noncompetitor pairs of firms. We benchmark the predictive models that use online metrics against those that use offline metrics. The offline metrics provide a strong benchmark, but we find that online metrics are competitive with or better than this benchmark. Also, we find that combining the online and offline metrics outperforms their individual usage.

We consider other nuances of competitor relationships by exploring predictive models of asymmetry in such relationships. We also study in detail the role of industry with respect to the competitor identification problem. We find that the predictive models are adept at discriminating between same (industrial) division competitors and noncompetitors. We consider the firm-ranking problem where all of the firms within the industrial division of a focal firm are ranked based on the likelihood of being a competitor to the focal firm. We find that the predictive models, on average, provide a more than five times higher return on effort (for managers) compared to the baseline rate while covering a fraction (10%) of the firms in the industrial division. This is illustrative of the benefit of complementing manual efforts in competitor identification with the suggested predictive model based on Web metrics of online isomorphism. We also uncover the utility of online isomorphism in providing leading indicators of new future competitors.

The proposed metrics and methods can be easily implemented as a SaaS (software as a service) platform, and the cost can then be amortized over many different managers and analysts (possibly across firms), and over time. As a future direction, with the help of domain experts, we would also like to validate how well our predictive models identify indirect or potential competitors. Currently, our metrics do not try to identify specific types of pages on a firm's website such as those describing products or partners. Automatic identification of such pages using heuristics or pattern recognition could provide additional value for competitor identification. Also, there is a potential for exploiting stylistic variables from websites.

As is popular in sociology and management literature, we consider competition at the level of firms, which is an important and relevant unit of analysis (Hannan and Freeman 1977). Firm-level analysis is also regularly used for purposes such as investments. However, if relevant gold standard data are available, it will be interesting to extend the study to other levels of granularity (supply-side versus demand-side, brands, etc.) and thus identify other subtleties of competitor relationships.

Firms are leaving increasingly large footprints on the Web that are indicative of their activities and relationships. We show that it is possible to harness different types of Web footprints of firms to measure their online isomorphism and use these measurements to build a predictive model of competitor relationships. The suggested Web metrics and the resulting decision tree models exploit the fact that firms, purposefully or inadvertently, step on the toes of their competitors, and their Web footprints, if carefully obtained, will reveal the same.

## Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/isre.2014.0563.

## References

Albert TC, Goes PB, Gupta A (2004) Gist: A model for design and management of content and interactivity of customer-centric web sites. *MIS Quart.* 28(2):161–182.

Bao S, Li R, Yu Y, Cao Y (2008) Competitor mining with the Web. *IEEE Trans. Knowledge Data Engrg.* 20(10):1297–1310.

Bar-Ilan J (2008) Informetrics at the beginning of the 21st century—A review. *J. Informetrics* 2(1):1–52.

Bergen M, Peteraf MA (2002) Competitor identification and competitor analysis: A broad-based managerial approach. *Managerial Decision Econom.* 23(4–5):157–169.

Bernstein A, Clearwater S, Provost F (2003) The relational vector-space model and industry classification. *Proc. 19th Internat. Joint Conf. Artificial Intelligence (IJCIA), Workshop Learn. Statist. Models from Relational Data, Acapulco, Mexico.*

Bernstein A, Clearwater S, Hill S, Perlich C, Provost F (2002) Discovering knowledge from relational data extracted from business news. *Proc. ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining, KDD 2002 Workshop on Multi-Relational Data Mining, Edmonton, Alberta.*

Chen MJ (1996) Competitor analysis and interfirm rivalry: Toward a theoretical integration. *Acad. Management Rev.* 21(1):100–134.

Clark BH, Montgomery DB (1999) Managerial identification of competitors. *J. Marketing* 63(3):67–83.

DiMaggio PJ, Powell WW (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *Amer. Sociol. Rev.* 48(2):147–160.

Egghe L, Rousseau R (1990) *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science* (Elsevier, Amsterdam).

Esrocka SL, Leichtya GB (2000) Organization of corporate Web pages: Publics and functions. *Public Relations Rev.* 26(3):327–344.

Flanagin AJ (2000) Social pressures on organizational website adoption. *Human Comm. Res.* 26(4):618–646.

Gartner (2013) Gartner survey. http://www.gartner.com/newsroom/id/2368315.

Hannan MT, Freeman J (1977) The population ecology of organizations. *Amer. J. Sociology* 82(5):929–964.

Heinze N, Hu Q (2006) The evolution of corporate Web presence: A longitudinal study of large American companies. *Internat. J. Inform. Management* 26(4):313–325.

Hill LN, White C (2000) Public relations practitioners perception of the World Wide Web as a communications tool. *Public Relations Rev.* 26(1):31–51.

Hill S, Provost F (2003) The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations* 5(2):179–184.

Hughes AM (2000) *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable Customer-Based Marketing Program* (McGraw-Hill, New York).

Kent ML, Taylor M, White WJ (2003) The relationship between Web site design and organizational responsiveness to stakeholders. *Public Relations Rev.* 29(1):63–77.

Lammers JC, Barbour JB (2006) An institutional theory of organizational communication. *Comm. Theory* 16(3):356–377.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Ma Z, Pant G, Sheng OR (2011) Mining competitor relationships from online news: A network-based approach. *Electronic Commerce Res. Appl.* 10(4):418–427.

Ma Z, Sheng OR, Pant G (2009) Discovering company revenue relations from news: A network approach. *Decision Support Systems* 47(4):408–414.

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* (Cambridge University Press, New York).

Mizruchi MS, Fein LC (1999) The social construction of organizational knowledge: A study of the uses of coercive, mimetic, and normative isomorphism. *Admin. Sci. Quart.* 44(4):653–683.

Padmanabhan B, Zheng Z, Kimbrough SO (2006) An empirical analysis of the value of complete information for eCRM models. *MIS Quart.* 30(2):247–267.

Pant G, Menczer F (2003) Topical crawling for business intelligence. *Proc. 7th Eur. Conf. Res. Advanced Tech. Digital Libraries* (*ECDL*) (Springer-Verlag, Berlin Heidelberg), 233–244.

Pant G, Srinivasan P, Menczer F (2004) Crawling the Web. Levene M, Poulovassilis A, eds. *Web Dynamics: Adapting to Change in Content Size, Topology and Use* (Springer-Verlag, Berlin Heidelberg), 153–178.

Peteraf MA, Bergen ME (2003) Scanning dynamic competitive landscapes: A market-based and resource-based framework. *Strategic Management J.* 24(10):1027–1041.

Porac JF, Thomas H (1990) Taxonomic mental models in competitor definition. *Acad. Management Rev.* 15(2):224–240.

Rivkin JW, Cullen A (2008) Finding information for industry analysis. Harvard Business School Background Note 708-481.

Robertson S (2004) Understanding inverse document frequency: On theoretical arguments for IDF. *J. Documentation* 60(5):503–520.

Russell Investments (2009) Russell 3000 index. http://www.russell.com/Indexes/data/fact_sheets/us/Russell_3000_Index.asp.

Santos BLD, Peffers K (1998) Competitor and vendor influence on the adoption of innovative applications in electronic commerce. *MIS Quart.* 34(3):175–184.

Scott J (1991) *Social Network Analysis: A Handbook* (Sage Publications, London).

Shannon CE (1948) A mathematical theory of communication. *Bell System Tech. J.* 27(3):379–423.

Shmueli G (2010) To explain or to predict? *Statist. Sci.* 25(3):289–310.

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Amer. Soc. Inform. Sci.* 24(4):265–269.

Turney P (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proc. Twelfth Eur. Conf. Machine Learn.* (Springer-Verlag, Berlin Heidelberg), 491–502.

Vaughan L, Gao Y (2006) Why are hyperlinks to business websites created? A content analysis. *Scientometrics* 67(2):291–300.

Walker BA, Kapelianis D, Hutt MD (2005) Competitive cognition. *MIT Sloan Management Rev.* 46(4):10–12.

Weiss GM, Provost F (2003) Learning when training data are costly: The effect of class distribution on tree induction. *J. Artificial Intelligence Res.* 19(1):315–354.

Witten IH, Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. (Morgan Kaufmann, San Francisco).

Zajac EJ, Bazerman MH (1991) Blind spots in industry and competitor analysis: Implications of interfirm (mis)perceptions for strategic decisions. *Acad. Management Rev.* 16(1):37–56.

Zander U, Kogut B (1995) Knowledge and the speed of the transfer and imitation of organizational capabilities: An empirical test. *Organ. Sci.* 6(1):76–92.

Zheng Z, Fader P, Padmanabhan B (2012) From business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data. *Inform. Systems Res.* 23(3–1):698–720.